Data Analysis and Visualization in R (IN2339)

Exercise Session 8 - Statistical Testing II

Hasan Celik, Christian Mertes, Vicente Yépez

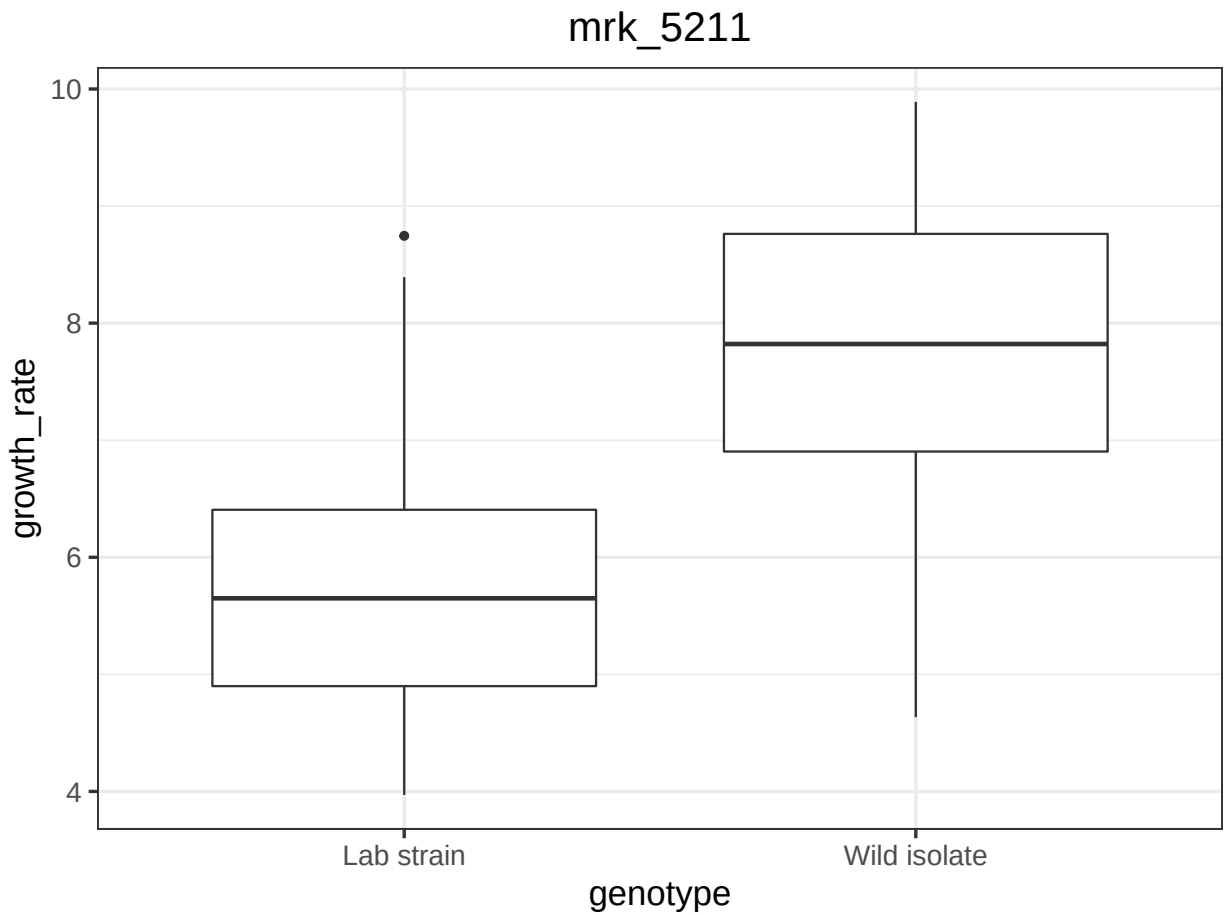
Section 00 - Getting Ready

1. Load yeast data and required packages

```
library(ggplot2)
library(data.table)
library(magrittr)
library(tidyr)
library(dplyr)
library(BBmisc)
gene <- fread("./extdata/eqtl/gene.txt")
genotype <- fread("./extdata/eqtl/genotype.txt")
genotype <- melt(genotype, id.vars = 'strain', variable.name = 'marker',
                 value.name = 'genotype')
growth <- fread("./extdata/eqtl/growth.txt")
growth <- melt(growth, id.vars = "strain", variable.name = 'media',
               value.name = 'growth_rate')
marker <- fread("./extdata/eqtl/marker.txt")
```

2. We reproduce the growth boxplot for each marker

```
getMaltoseDt = function(mrk){
  growth_mrk <-  merge(growth, genotype[marker == mrk, .(strain, genotype)],
                       by = 'strain')
  growth_mrk[media == "YPMalt"]
}

# boxplot
plot_growth_one_mk <- function(mk){
    ggplot(getMaltoseDt(mk), aes(genotype, growth_rate)) +
    geom_boxplot() +
    labs(title = mk) + theme_bw(base_size = 16) +
    theme(plot.title = element_text(hjust = 0.5))
}
plot_growth_one_mk("mrk_5211")
```



Section 01 - Test the association between markers and growth

1. Testing marker 5211

Last week, using permutation, we saw that some markers associated with growth. Which of the statistical tests from the lecture would you use to test this association? For each marker, we are not certain if the genotype will cause a positive or negative effect on growth; therefore, which kind of alternative hypothesis would you choose: double-sided, right or left? Apply the test to marker 5211 to obtain a p-value and see if the association that we found last week still holds. If more than one of the tests are appropriate, use them all and compare the results.

```
## We can use either Wilcoxon or the t-test (with or without Welch's correction)

m_dt <- getMaltoseDt('mrk_5211')

## Because the growth can be affected in both directions,
## we choose double-sided (default).

## Student's t-Test with Welch correction
### NB: in this case, using an equal variance t-test will give indistinguishable results
### But generally, using the Welch test is usually the better bet
### As variances are unlikely to be equal for most interesting comparisons
tt <- t.test(alternative="two.sided", growth_rate ~ genotype, m_dt)
```

```
### NB: instead of the formula, one can also supply the groups manually:
#tt <- t.test(alternative="two.sided",
#    m_dt[genotype == 'Lab strain', growth_rate],
#    m_dt[genotype == 'Wild isolate', growth_rate])
tt
```

```
##
##  Welch Two Sample t-test
##
## data:  growth_rate by genotype
## t = -10.805, df = 152, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.405189 -1.661599
## sample estimates:
##   mean in group Lab strain mean in group Wild isolate
##                   5.763086                   7.796480
```

```
## Wilcoxon Rank Sum Test
w <- wilcox.test(alternative="two.sided", growth_rate ~ genotype, m_dt)
w
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  growth_rate by genotype
## W = 690, p-value = 2.264e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
## Both tests give the same results and are in accordance
## with what we found last week.
```

2. Applying the test to other markers

Make a function that given a marker and the name of a statistical test, returns the p-value of the association of that marker wrt growth. Test your function on the markers 1653 and 5091. Since we used the same marker last week: do you get similar results as last week?

```
test_growth <- function(mk, test=c('wilcoxon', 't')){

  # Only wilcoxon or t.tests allowed or the abbreviations
  test <- match.arg(test)

  m_dt <- getMaltoseDt(mk)

  if(test == 'wilcoxon') {
          pval <- wilcox.test(alternative="two.sided", growth_rate ~ genotype, m_dt)$p.value
    } else {
          pval <- t.test(alternative="two.sided", growth_rate ~ genotype, m_dt)$p.value
  }
  return(pval)
}
# A solution using switch(test, ..) is more elegant


test_growth('mrk_1653', test = 'w')
```

```
## [1] 0.5009288
```

```
test_growth('mrk_1653', test = 't')
```

```
## [1] 0.4971713
```

```
test_growth('mrk_5091', test = 'wilcox')
```

```
## [1] 0.000808814
```

```
test_growth('mrk_5091', test = 't')
```

```
## [1] 0.0006307451
```

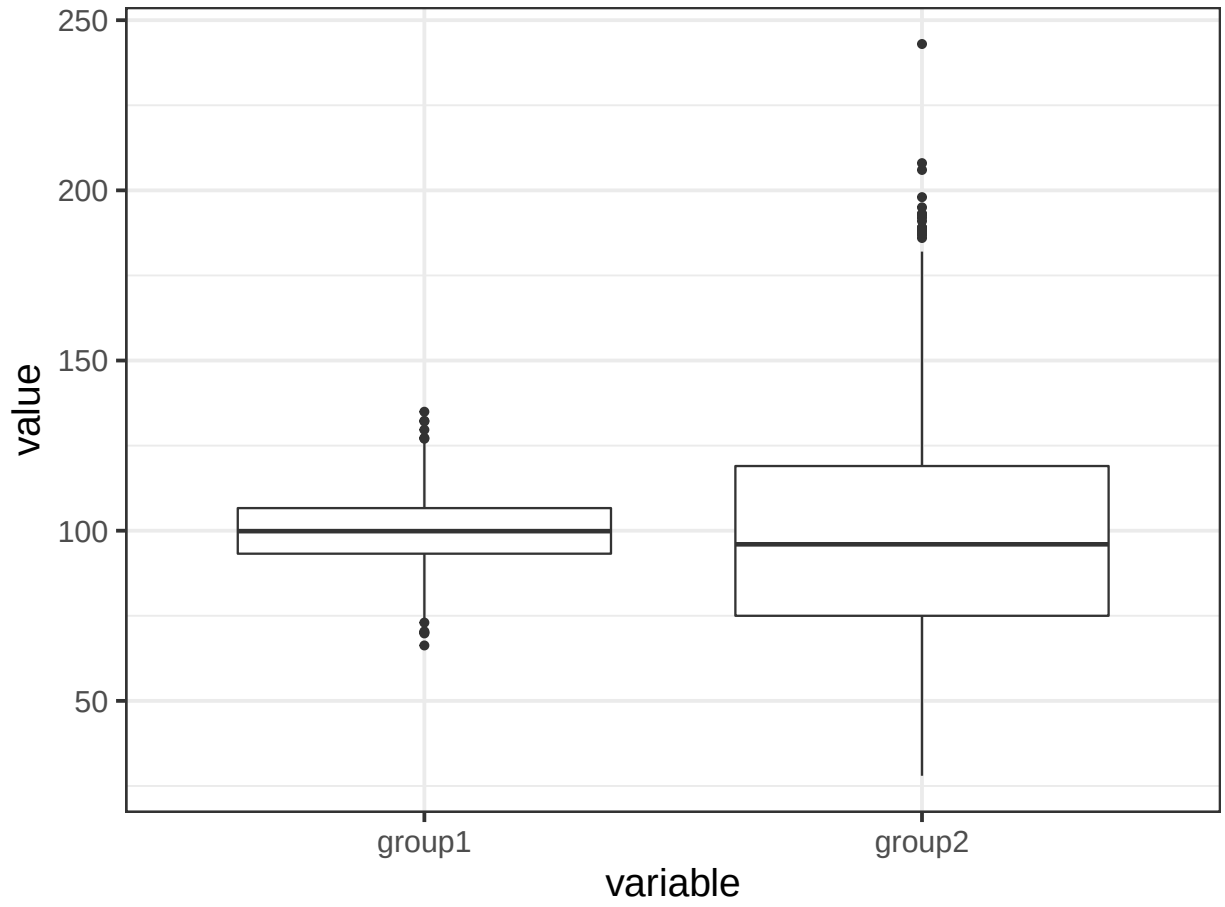
```
## We conclude the same as before:
## marker 1653 is not associated with growth, but marker 5091 is.
```

Section 02 - Pitfalls when using the Student's t-Test

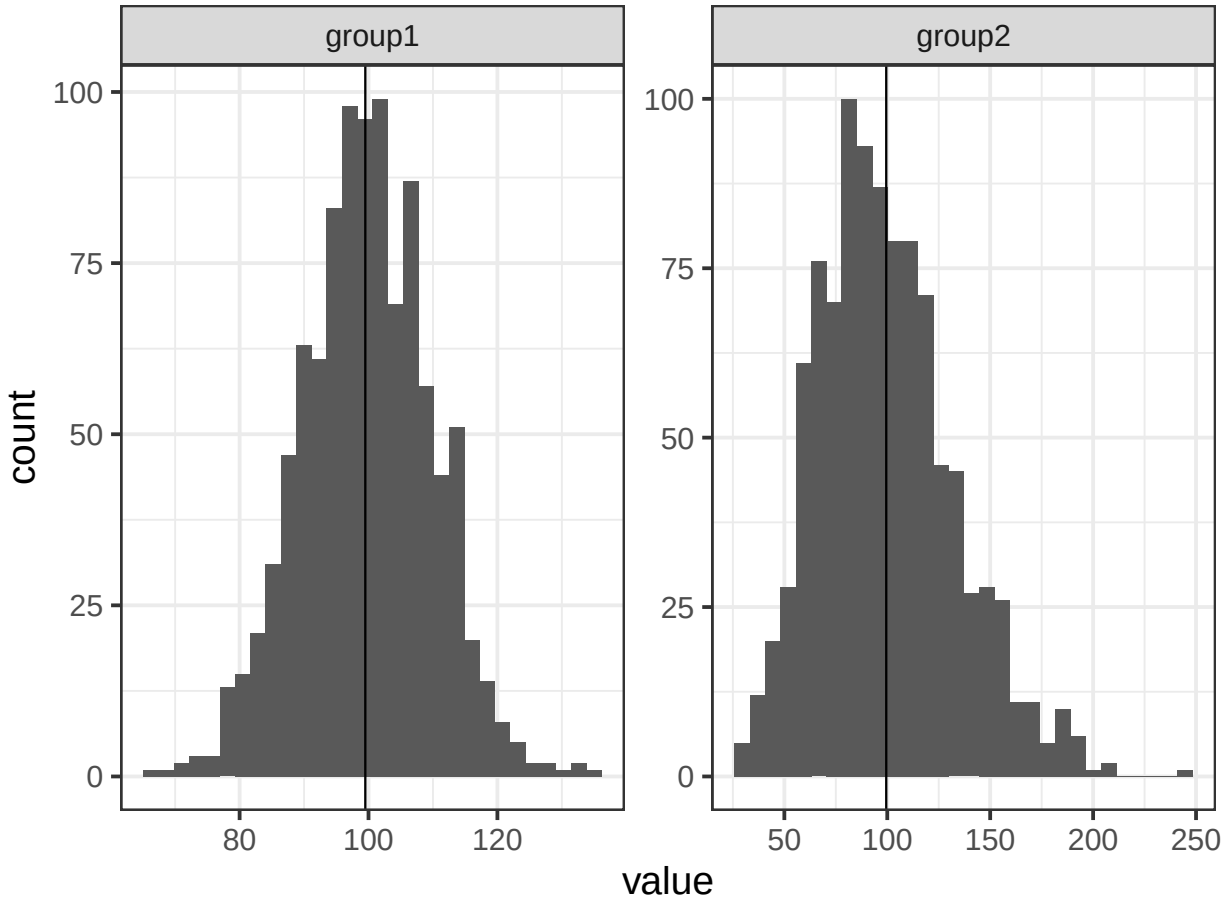
1. Load the data from the **stats-pitfalls.csv** file and visualize the data. Apply both the t-test and Wilcoxon test on it. What do you observe? Which test is the better choice here?

Hint: use stat_summary() or geom_vline() to add points/lines to the ggplot object.

```
## load the data
dt <- fread("extdata/stats-pitfalls.csv")
x <- dt$group1
y <- dt$group2

## Use the boxplot to compare two groups
ggplot(melt(dt), aes(variable, value)) + geom_boxplot() +
  theme_bw(base_size = 18)
```




```
## Use the histogram to investigate the different distributions per group
ggplot(melt(dt), aes(value)) + geom_histogram() +
  facet_wrap(~variable, scales = 'free') +
  geom_vline(aes(xintercept=mean(value))) +
  theme_bw(base_size = 18)
```



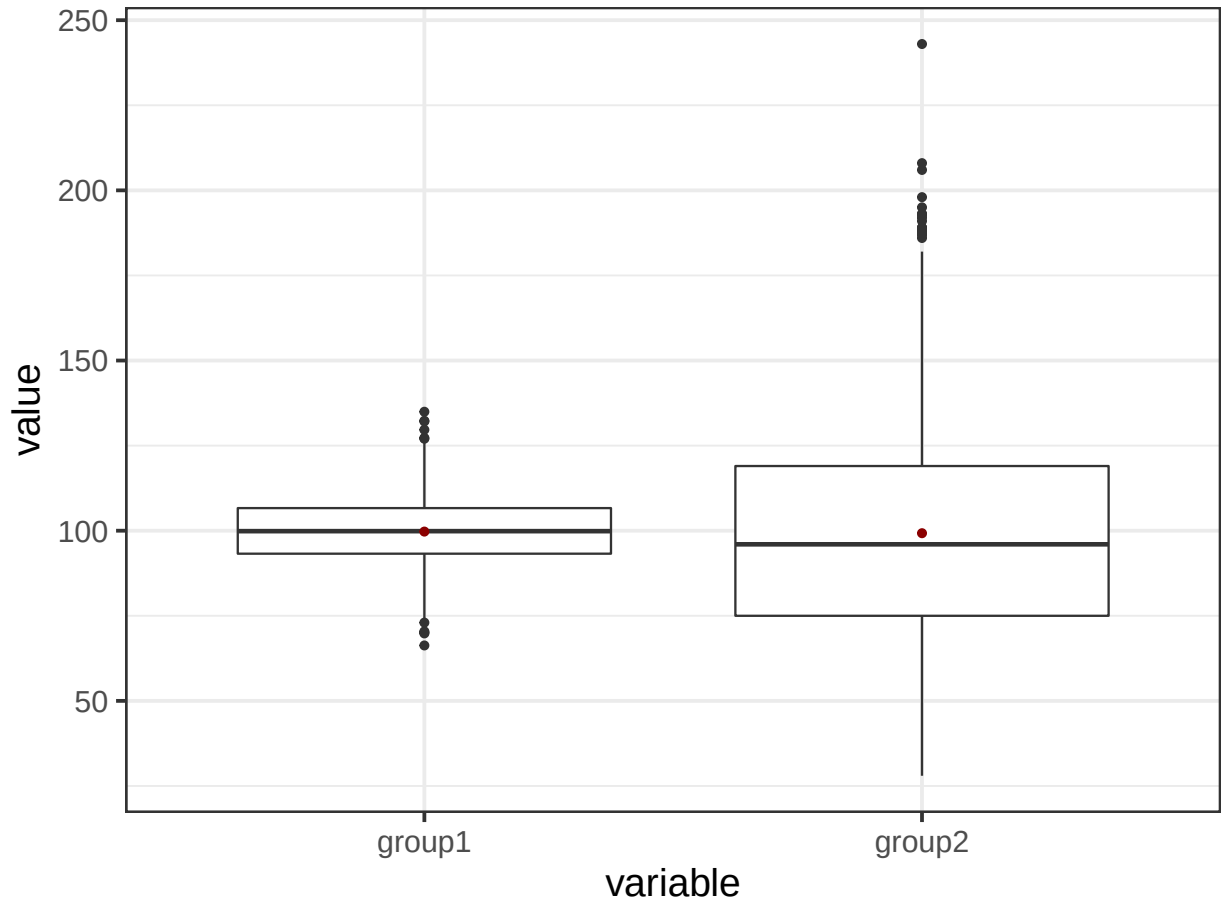

```
# The second group doesn't seem to follow a Normal distribution

## Apply the Wilcoxon test and the t-test
wilcox.test(x,y)$p.value    # significant
```

```
## [1] 0.001594905
```

```
t.test(x,y)$p.value         # not significant
```

```
## [1] 0.6625758
```

```
# Which one is correct?

## The boxplot uses the median and not the mean. Therefore we add the mean to it.
ggplot(melt(dt), aes(variable, value)) + geom_boxplot() +
  stat_summary(fun.y=mean, geom="point", col="darkred") +
  theme_bw(base_size = 18)
```




```
## The t-test only compares the mean of each group and since the
## second group is not gaussian distributed, the t-Test fails to
## detect the differences. With a non gaussian distribution the
## assumptions of the t-Test are violated and hence the
## Wilcoxon test is the right choice in this situation.
```

Section 03 - Test the association between 2 markers

1. Last week, using permutation, we saw that marker 5091 is significantly associated with marker 5211. Which of the statistical tests from the lecture would you use to test this association? Apply it to obtain a p-value and see if the association still holds. Note: association can mean that 2 markers choose the same genotype more often than usual, or that 2 markers choose different genotype more often than usual.

Use the contingency table:

```
mks_geno <- genotype[marker %in% c('mrk_5091', 'mrk_5211')] %>%
  spread(marker, genotype)
table(mks_geno[, 2:3])
```

```
##                mrk_5211
## mrk_5091        Lab strain Wild isolate
##   Lab strain            53           27
##   Wild isolate          25           53
```

```r
# Ho: Marker 5091 is not significantly associated with marker 5211
## The appropriate test is a Fisher's Exact test
fisher.test(table(mks_geno[, 2:3]))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table(mks_geno[, 2:3])
## p-value = 1.821e-05
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  2.037578 8.542156
## sample estimates:
## odds ratio
##   4.120873
```

```r
# p-value = 1.82e-5, which is significant
## The alternative 'greater' implies that both markers choose lab
## and wild together more often than usual.
## The alternative 'less' implies that if one marker chooses lab,
## then the other one would choose 'wild', more often than usual.
## The alternative 'double.sided' would test for both events.
```

2. Make a function that given any two markers, returns the p-value of the appropriate statistical test to evaluate the association between the markers. Give the alternative as a parameter of the function. Test your function for, e.g., mrk_1 vs mrk_13314.

```r
marker_test <- function(marker1, marker2, alternative = 'two.sided'){
  mks_geno <- genotype[marker %in% c(marker1, marker2)] %>%
    spread(marker, genotype)

  table_markers <- table(mks_geno[, 2:3])
  pval <- fisher.test(table_markers, alternative = alternative)$p.value
  return(pval)
}
marker_test('mrk_1','mrk_13314', alternative = 'two.sided')
```

```
## [1] 0.7528492
```

Section 04 - Test the association between markers and genotype

1. We know that each marker is composed from 2 different genotypes: lab and wild strains. We assumed that the ratio between them is 1:1 (or 50%-50%). Compute the actual ratio.

```r
# Let's take the lab strain as reference
genotype[genotype == 'Lab strain', .N] / nrow(genotype)
```

```
## [1] 0.503981
```

```r
# Close enough to 0.5. For simplicity, we will assume 0.5 for the next exercises.
```

2. We are interested in testing whether a marker is associated with a genotype, i.e., if a marker has more lab strain or wild isolate than the expected 50%. Which of the statistical tests from the lecture would you use to test this association? Apply it to markers 3385 and 13314 to obtain a corresponding p-value. Make a function out of it.

```
# Ho: markers take lab or wild genotype
# with 50% chance
## The right test is the Binomial Test

mk_geno <- genotype[marker == 'mrk_13314']
binom.test(mk_geno[genotype == 'Lab strain', .N], nrow(mk_geno),
           p = .5, alternative = 'two.sided')$p.value
```

```
## [1] 0.5777428
```

```
# Because it's a two.sided test with p = 0.5,
# it doesn't matter if we compare the # of lab strain or of wild isolate
binom.test(mk_geno[genotype == 'Wild isolate', .N], nrow(mk_geno),
           p = .5, alternative = 'two.sided')$p.value
```

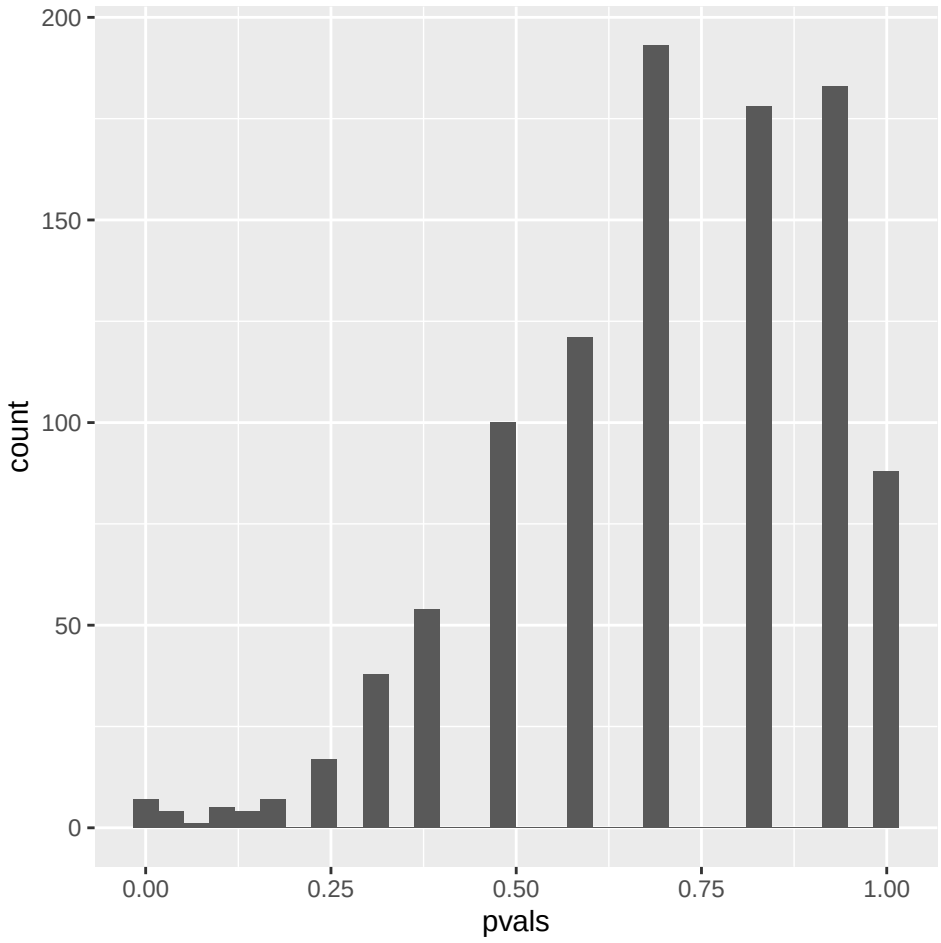
```
## [1] 0.5777428
```

```
# function
genotype_marker_test <- function(mk){
  mk_geno <- genotype[marker == mk]
  binom.test(mk_geno[genotype == 'Lab strain', .N], nrow(mk_geno),
             p = .5, alternative = 'two.sided')$p.value
}

genotype_marker_test(mk = 'mrk_3385')
```

```
## [1] 0.0005801271
```

3. So far, we have hand-picked some markers. Apply the corresponding test to all markers and obtain a distribution of p-values. Make a histogram out of that distribution. Further, compute how often we reject the null hypothesis that the ratio is 0.5 at the 5% level. What do you observe?

```
markers <- unique(genotype$marker)
pvals <- sapply(markers, function(mk){
  genotype_marker_test(mk)
  })
names(pvals) <- markers

ggplot(data.table(pvals=pvals), aes(pvals)) + geom_histogram()
```




```
sum(pvals < 0.05)/length(pvals)
```

```
## [1] 0.011
```

```
# Even if the null hypothesis is always true
# We would expect to reject the null hypothesis around 5% of the time
# We see that we reject the null hypothesis considerably less frequently than 5%
# This means that the genotype configuration is stricter than a random 50%-50%.
# In other words, we are more concentrated around the 50:50 point
# Than a binomial distribution would expect, likely as a result of the
# experimental design.
```

Section 05 - Correlations and correlation tests

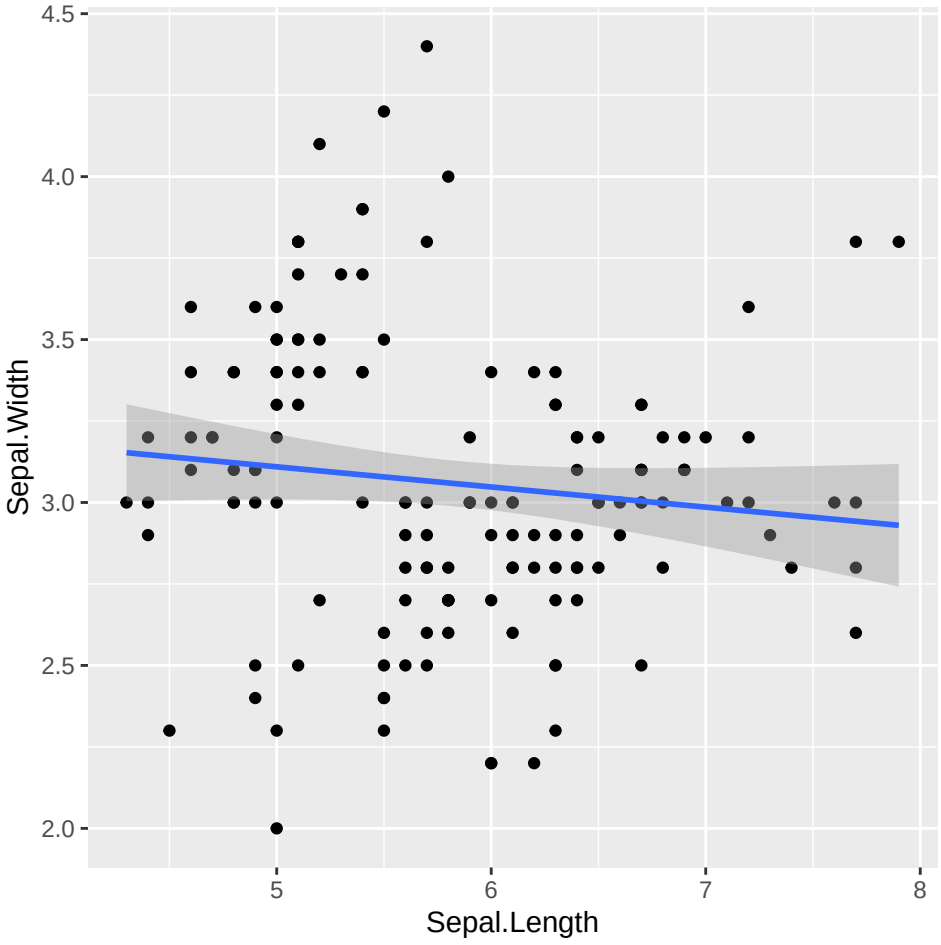
1. Investigate the correlation between `Sepal.Length` and `Sepal.Width` within the iris dataset. Calculate the correlation and plot your results. Are there any issues with your results? Discuss.

```
library(datasets)
data(iris)

cor_value = cor.test(iris$Sepal.Length, iris$Sepal.Width, method = "pearson")
cor_value
```

```
##
```

```
##  Pearson's product-moment correlation
##
## data:  iris$Sepal.Length and iris$Sepal.Width
## t = -1.4403, df = 148, p-value = 0.1519
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.27269325  0.04351158
## sample estimates:
##        cor
## -0.1175698
```

```
# cor_value = cor.test(iris$Sepal.Length, iris$Sepal.Width, method = "spearman")
# cor_value

ggplot(iris, aes(x=Sepal.Length, y=Sepal.Width)) +
  geom_point() +
  geom_smooth(method=lm)
```




```
# + geom_label(x = 7.5, y = 4.25, label = paste('Cor = ', round(cor_value$estimate, 2)))

# We obtain a negative correlation!!??
## The data can depend on multiple variables, i.e different plant species.
## In this case the correlation of the data is confounded. We could remove the
## confounding effects with PCA and then run the correlation analysis
```

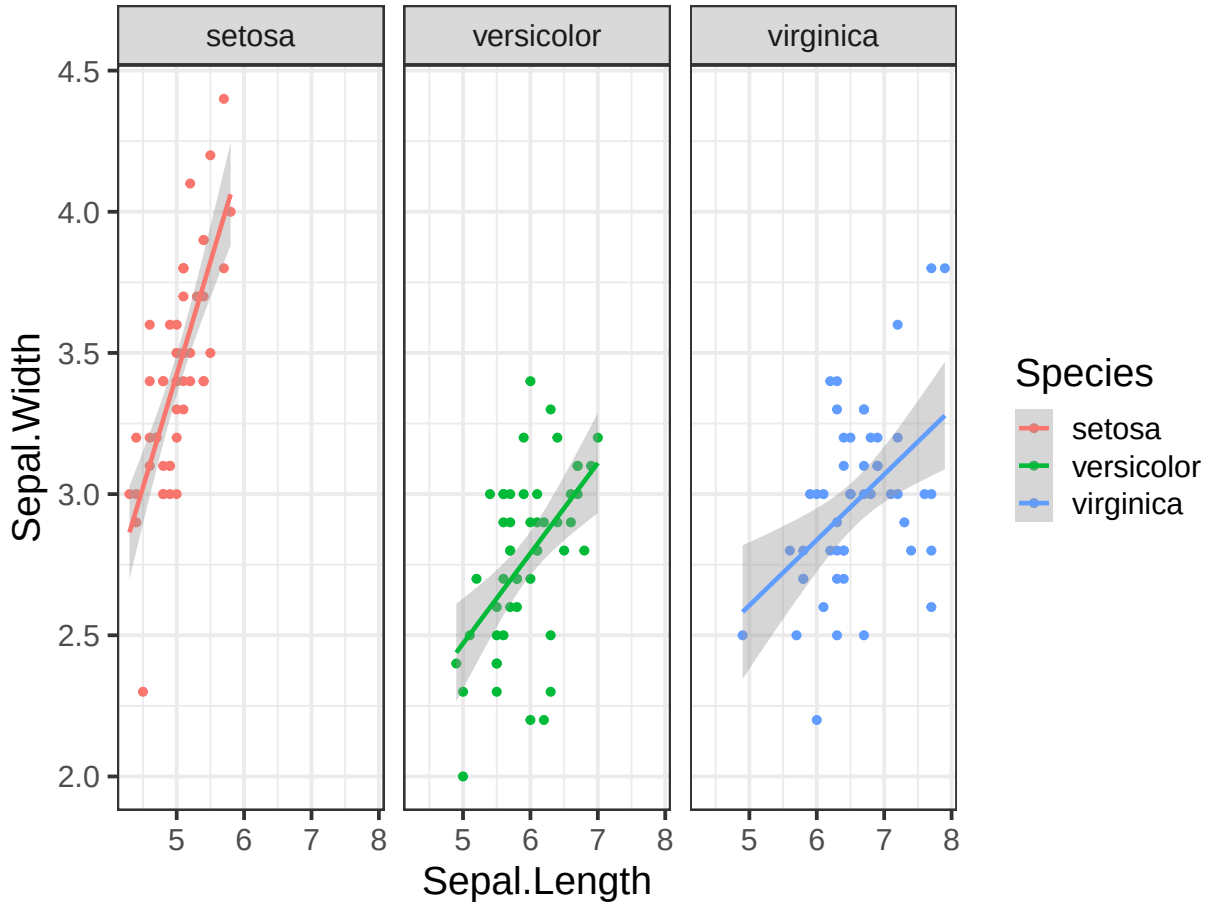
```
## or run the analysis in the univariate case for each group.
```

2. Repeat the previous analysis for each species independently. How does the correlation between `Sepal.Length` and `Sepal.Width` change?

```
iris_dt <- as.data.table(iris)
corr_dt <- iris_dt[, cor.test(Sepal.Length, Sepal.Width, method="pearson"),
                   by = Species]
corr_dt
```

```
##       Species statistic parameter      p.value  estimate null.value alternative
## 1:     setosa  7.680738        48 6.709843e-10 0.7425467          0   two.sided
## 2:     setosa  7.680738        48 6.709843e-10 0.7425467          0   two.sided
## 3: versicolor  4.283887        48 8.771860e-05 0.5259107          0   two.sided
## 4: versicolor  4.283887        48 8.771860e-05 0.5259107          0   two.sided
## 5:  virginica  3.561892        48 8.434625e-04 0.4572278          0   two.sided
## 6:  virginica  3.561892        48 8.434625e-04 0.4572278          0   two.sided
##                                  method                    data.name  conf.int
## 1: Pearson's product-moment correlation Sepal.Length and Sepal.Width 0.5851391
## 2: Pearson's product-moment correlation Sepal.Length and Sepal.Width 0.8460314
## 3: Pearson's product-moment correlation Sepal.Length and Sepal.Width 0.2900175
## 4: Pearson's product-moment correlation Sepal.Length and Sepal.Width 0.7015599
## 5: Pearson's product-moment correlation Sepal.Length and Sepal.Width 0.2049657
## 6: Pearson's product-moment correlation Sepal.Length and Sepal.Width 0.6525292
```

```
ggplot(iris, aes(x=Sepal.Length, y=Sepal.Width, color=Species)) +
  geom_point() +
  geom_smooth(method=lm) +
#  geom_label(data = corr_dt, mapping = aes(x = 6.5, y = 4.2,
#    label=paste('Cor = ', round(estimate, 2)))) +
  facet_grid(. ~ Species) + theme_bw(base_size = 18)
```




```
## Since the different species are treated independently the correlation is
## not confounded anymore.
```

3. We want to know whether there is a correlation between attendance to the exercise sessions and the points achieved in the final exame of the students in our course based on the previous years' data. We provide simulated data below (due to legal restrictions). Load the data from **exam_correlation.tsv**. Calculate the correlation between attendance and points using `Pearson` and `Spearman` methods and visualize it. Some students will drop out of the distribution since they were planning to take the retake exam and skipped the first exam, thus obtaining a grade of zero. Which correlation method should be preferred in this context and why?

```
ex_dt <- fread("extdata/exam_correlation.tsv")

# compute the correlation
per_cor <- ex_dt[,cor.test(attendance, achieved_points, method="pearson")]
per_cor
```

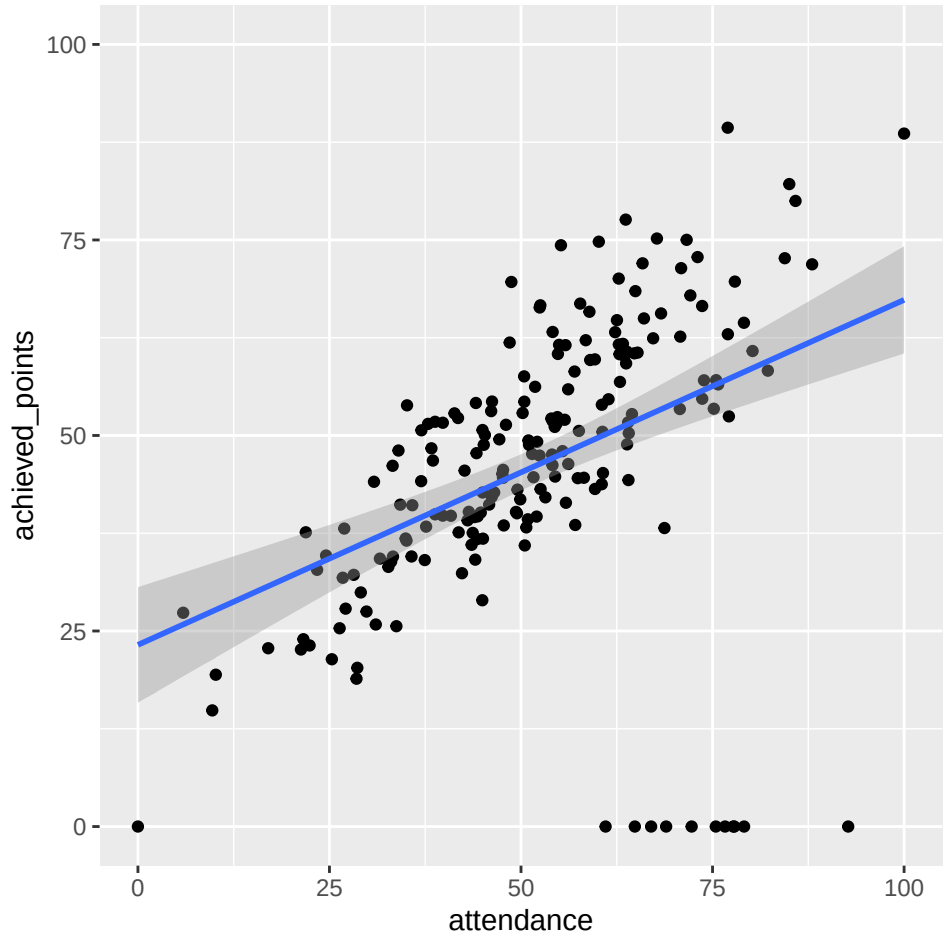
```
##
##  Pearson's product-moment correlation
##
## data:  attendance and achieved_points
## t = 6.4551, df = 198, p-value = 8.159e-10
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2953046 0.5253140
```

```
## sample estimates:
##       cor
## 0.4169623
```

```
spr_cor <- ex_dt[,cor.test(attendance, achieved_points, method="spearman")]
spr_cor
```

```
##
##  Spearman's rank correlation rho
##
## data:  attendance and achieved_points
## S = 578278, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.5662806
```

```
# visualize the data and the correlation
ggplot(ex_dt, aes(x=attendance, y=achieved_points)) +
    geom_point() +
    geom_smooth(method=lm) +
#    geom_label(x = 20, y = 80, label = paste('Pearson = ', round(per_cor$estimate, 2))) +
#    geom_label(x = 20, y = 90, label = paste('Spearman = ', round(spr_cor$estimate, 2))) +
    ylim(0, 100)
```



```
## Since we do see some outliers in our dataset, the Spearman method should be used.
## The Spearman method is robust against outliers within the dataset.
```